Case Study: Ames Iowa Housing Prices

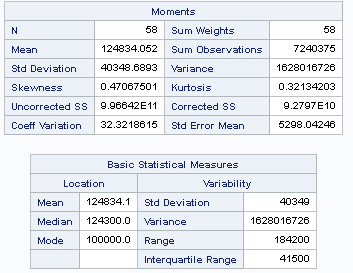
# Data Description:

Training set – Target Variable Characteristics:

# 

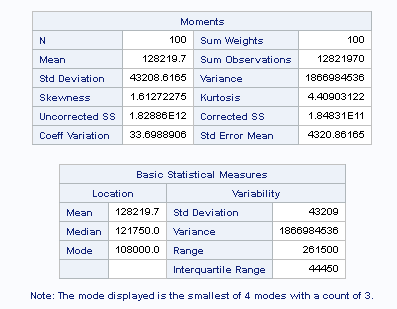
The training set consists of 383 observations with an average SalePrice ~= 138k.

For the BrkSide neighborhood specifically:



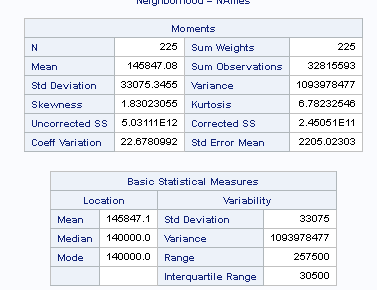
The training set consists of 58 observations with an average SalePrice ~= 125k.

For the Edwards neighborhood specifically:



The training set consists of 100 observations with an average SalePrice ~= 128k.

For the Edwards neighborhood specifically:



The training set consists of 225 observations with an average SalePrice ~= 145k.

### Train/Test Balance:

Data seems well balanced by neighborhood with respect to the explanatory variables in QOI #1 (GrLivArea and Neighborhood)

|  |  |
| --- | --- |
|  |  |
|  |  | **Count** | | | **Average Square Footage** | | |
| **Neighborhood** | **Average Sale Price** | **Train** | **Test** | **Balance** | **Train** | **Test** | **Balance** |
| BrkSide | 124,834 | 58 | 50 | 86% | 1,203 | 1,272 | 106% |
| NAmes | 145,847 | 225 | 218 | 97% | 1,310 | 1,273 | 97% |
| Edwards | 128,220 | 100 | 94 | 94% | 1,340 | 1,335 | 100% |

Ultimately, training the model will be done on a 5(?) fold cross validation scheme and then the model will be run against a test set

# Problem Statement:

Century 21 seeks to answer two questions related to housing prices in Iowa:

* QOI 1: Is there a relationship between square footage and sale price?
* QOI 2: Is this relationship dependent on neighborhood?

The data set we will be using will only consider a subset of the available features we could use to build a better model. Since the analysis at hand only seeks to answer questions related to house size and location – we will limit the analysis to only those variables and leave further explanatory variable dependencies for future studies.

# Modeling:

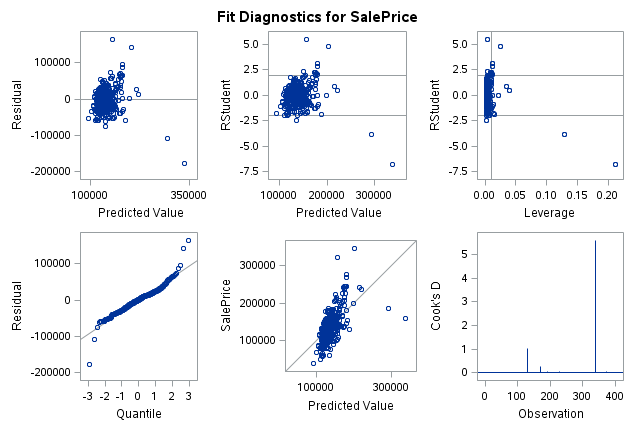
## QOI: Is there a relationship between square footage (GrLivArea) and sale price (SalePrice)

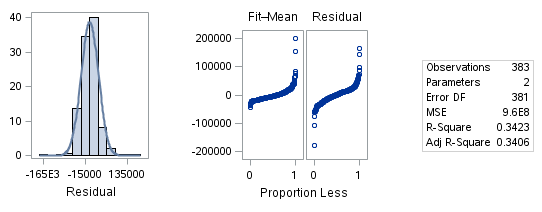
To answer this question we will first run a basic linear regression of the form:

SalePrice = 0 + 1 GrLivArea

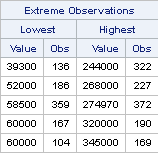
## Checking Assumptions

#### Normality:





Scatter, Q-Q plot and histogram of residuals are all relatively normal. There appear to be some influential observations for the following:



Observation 322 appears problematic and is a candidate for being dropped (cook’s D >5)

Let’s run the analysis with all of the extreme observations removed and see if there is a significant difference in the predictive power of the model:

NO OUTLIERS:

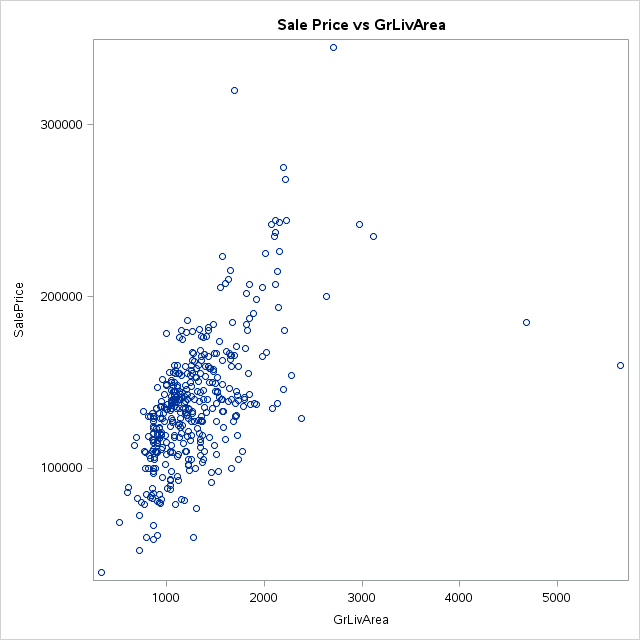


WITH OUTLIERS:



Influence of outliers appears to be minimal – we can check final model with/without them to assess affect on predictive capability of model. In building a larger model – we may want to keep them – as the addition of new features may explain the deviation in the norm.

#### Linear Trend:



The mean sale price as a function of GrLivArea does appear to be relatively linear based on the scatter plot above (increasing GrLivArea tends to have a positive effect on the mean sale price)

#### Equal SD:

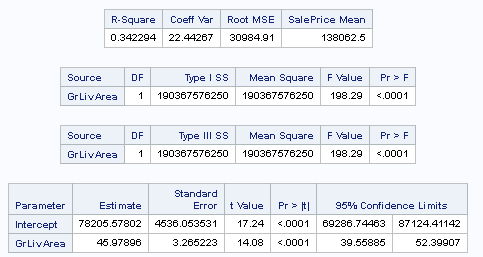
Within the prediction band, the SD of prices appears relatively constant (few signs of heteroscedasticity)

#### Independence:

No variables are assumed to be interdependent in this example.

**Model assumptions for basic linear regression seem to be well met. We will proceed with the analysis as is (inference on means)**

## Simple Linear Model Results:



Assuming the only variable that contributes to SalePrice to be GrLivArea – there appears to be strong evidence of a non-zero intercept (0 = {69,286 – 87,124}). In this instance the intercept can be thought of as the cost of the land (i.e a house with no square footage).

A strong direct relationship between GrLivArea and SalePrice also exists, accounting for an effect of 45.9$/sqft of living space – or to state it more clearly – an increase of approx 4,600$ in sale price can be expected from every increase of 100sqft of living space. A 95% CL puts this accretive affect at anywhere from 3,900/100sqft to 5,200/100sqft.

Given the low R-square value for this regression (0.34) it is clear that there are other variables which contribute to SalePrice. GrLivArea only accounts for about 34% of the variance in SalePrice.

## QOI 2: Does relationship between square footage (GrLivArea) and sale price (SalePrice) vary based on neighborhood?

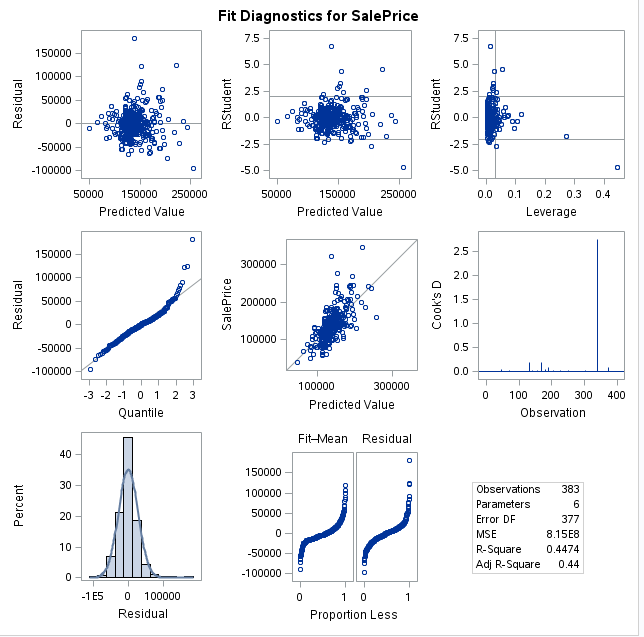
To address this question we will consider a regression of the form (including interaction terms)

SalePrice = 0 + 1 GrLivArea + 2 BrkSide + 3 Edwards + 4 GrLivArea|Edwards + 5 GrLivArea|BrkSide

\*reference level for categorical variable is NAmes

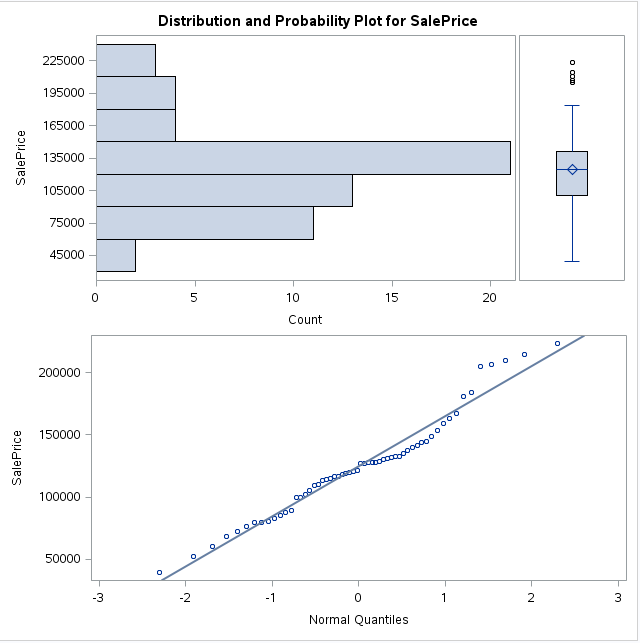
## Revisiting Assumptions:

#### Normality



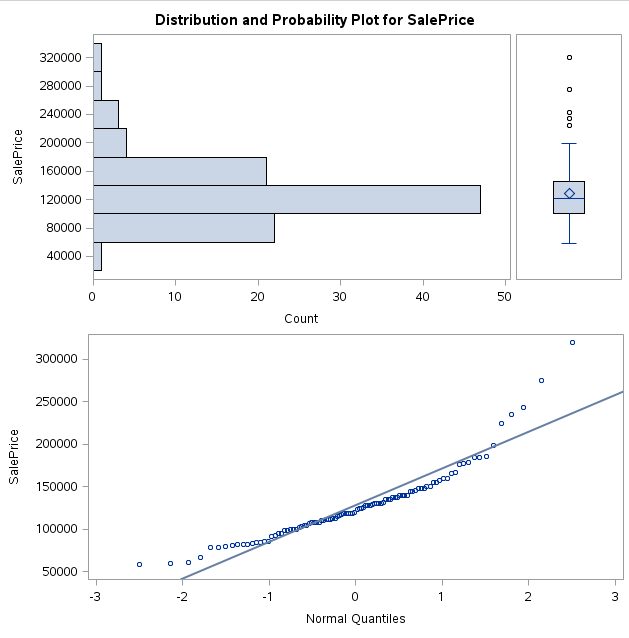
Final model may want to consider removal of high leverage observations – but data is otherwise clean.

##### BrkSide



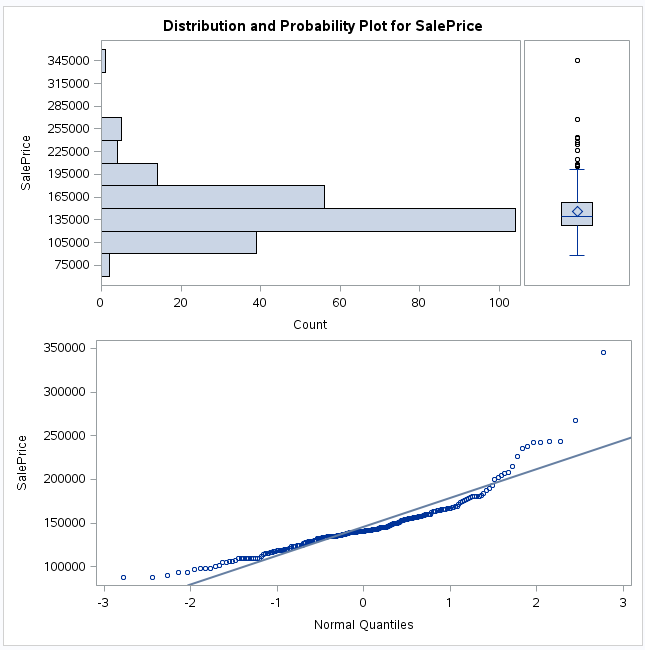
BrkSide shows strong evidence of positive skew though not enough to violate the assumption of normality.

##### Edwards



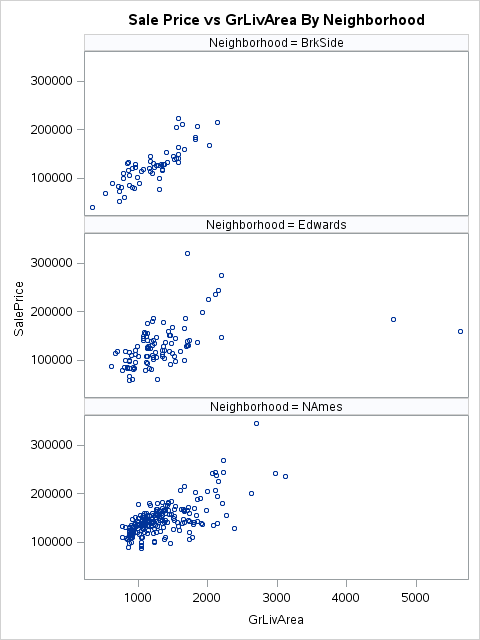
Edwards also shows strong evidence of positive skew though not enough to violate the assumption of normality.

##### NAmes



Of the three neighborhoods, Names exhibits the most positive skew though not enough to violate the assumption of normality.

#### Linear Trend



Generally speaking – each of the neighborhoods exhibits a linear relationship between GrLivArea and SalePrice. Edwards being somewhat suspect in that regard having observations with very high square footages and very low sale prices. Possible outliers , observations 524 and 1299 were both partial sales – this could be useful info down the road.

#### Equal SD

Scatter plots do not show any sign of heteroscedasticity amongst the 3 neighborhoods.

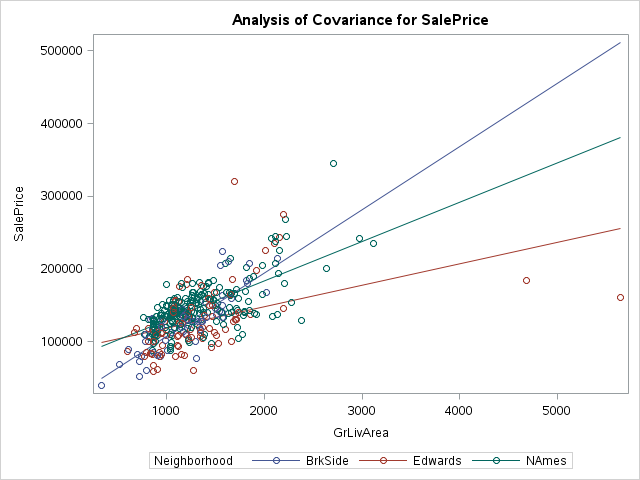
#### Independence

Since being in one neighborhood implicitly excludes a home from being in the others – we should be fine to assume independence.

## Linear Model With Neighborhood Results



Assuming an unequal slope model (interaction terms included):



Given the full model (5 parameters with interaction terms included we see the following relationships)

* Controlling for the effect of square footage, the average home in